DOCUMENT RESUME

ED 294 888 TM 011 411

AUTHOR Tallmadge, G. Kasten
TITLE How and How Much Do Biases Resulting from Regression to the Mean Affect Outcome Evaluations?
PUB DATE Apr 88
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Correlation; *Predictive Measurement; Pretests Posttests; *Regression (Statistics); Statistical Analysis; Statistical Bias
IDENTIFIERS *Bilingual Education Evaluation System; *Title I Evaluation and Reporting System

ABSTRACT

The question of regression to the mean is discussed in the contexts of the Title I Evaluation and Reporting System (TIERS) and the Bilingual Education Evaluation System (BEES), both of which involve assessing the achievement growth of project students from pre- to post-test. The correction formula recommended for use with TIERS was designed to eliminate regression between selection and pre-test; for BEES, the idea of a separate selection test was eliminated and an attempt was made to eliminate regression between the selection/pre-test and the post-test and to estimate total group correlation from the actual correlation in the selected sub-group. How much regression there is to correct for and how well alternative correction procedures work are discussed. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED294888

How and How Much Do Biases Resulting from Regression to the Mean Affect Outcome Evaluations?

G. Kasten Tallmadge

RMC Research Corporation

April 1988

Paper presented at the Annual Meeting of the
American Educational Research Association
New Orleans, Louisiana

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. KASTEN TALLMADGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 011 411

I was involved in developing both the Title I Evaluation and Reporting System called TIERS and the Bilingual Education Evaluation System called BEES. You will note that we left the reporting part out of the BEES acronym even though the system includes a reporting component. Our client, the Department of Education, thought that BEERS might be somewhat too frivolous.

Both TIERS and BEES involve assessing the achievement growth of project students from pre- to posttest. In our initial TIERS formulation, we cautioned users not to use the same set of test scores both to select the students who would participate in the project and, subsequently, as their pretest measures. We were aware that this practice would produce spuriously inflated growth estimates because of the phenomenon known as regression to the mean.

Our original recommendation was to select project participants first, using teacher judgments, test scores, a combination of the two, or any other indicators that were available or could be obtained. Once students were selected, they could be pretested. This practice would, we argued, minimize the bias due to regression to the mean.

While our recommendation was a sound one, we were immediately bombarded with complaints from the field that separate selection and pretests imposed an unacceptable testing burden on both schools and students. We should come up with some way that a single set of scores could be used for both purposes.

In theory, this task was not too difficult. It is well documented that in homoskedastic, bivariate normal distributions, the amount of regression to the mean that a selected subgroup will show can be predicted from (a) the distance between the mean score of the selected subgroup and that of the total group tested, and (b) the pretest-posttest correlation for the total group. The formula for the corrected pretest mean is:

$$X_s' = X_s + (1 - r_{xy})(X_g - X_s)$$

Unfortunately, homoskedastic, bivariate normal distributions are an endangered species--in fact, they appear to exist only in textbooks. Nevertheless, without very much knowledge about how well the formula would work with less well behaved distributions, we proceeded to offer the correction formula given above as a recommended approach for removing the regression-effect bias when the same test scores are used for both selection and pretesting purposes.

Using the formula required that the total group's pretest-posttest correlation be estimated since unselected students were rarely posttested. Even if they were, the total group's correlation could be significantly lowered by the fact that the selected subgroup received a special intervention.

The procedure we recommended for estimating r_{xy} begins with the test publisher's reliability coefficient. If the publisher provides a test-retest reliability coefficient, we recommended it be used. We developed constants to subtract from other types of reliability coefficients to obtain estimated test-retest reliabilities in case the publisher did not provide real ones.

The next step was to adjust the actual or estimated test-retest reliability coefficient for the restricted range of talent existing in the local group compared to the test publisher's nationally representative standardization sample. This was to be done following standard procedures, using the ratio of the two groups' variances.

We didn't get a lot of complaints about the technical merits of adjustment procedures. Local evaluators found it cumbersome, however, and a substantial number of them decided to ignore it--possibly because unadjusted growth estimates are always larger than adjusted ones.

The complaints we did hear from professional measurement and evaluation people--notably Gene Glass (1978) and Bob Linn (1982)--centered on the fact that our original recommendation to use separate selection and pretests did not totally eliminate the regression problem. The argument is as follows:

- Using separate selection and pretests eliminates regression from the selection test to the pretest.

- It also eliminates most of the regression from the selection test to the posttest.

- However, to the extent that the selection-posttest correlation is lower than the selection-pretest correlation, there will be additional regression from pre- to posttest.

Several investigators (Linn, 1980; Roberts, 1980; Tallmadge, 1982) have examined this issue and seem to agree that the additional regression from pre- to posttest amounts to

approximately 1 NCE for typical Chapter 1 projects serving students above second grade. For grades K, 1, and 2, it appears that the additional regression amounts to about 2 NCEs, although, of course, it will vary somewhat as a function of the selection-test cutoff score and the correlations among the selection, pretest, and posttest scores.

The correction formula we recommended for use with TIERS was designed to eliminate regression between selection and pretest--thus bias due to any additional regression from pre- to posttest would remain, regardless of whether a separate selection test was administered or the recommended statistical adjustment was made.

With BEES, several of our recommendations changed. First of all, we yielded to public demand and abandoned the idea of a separate selection test. Then we decided to try to get rid of all the regression between the selection/pretest and the posttest. Finally, we decided to estimate the total-group correlation not from a reliability coefficient but from the actual correlation in the selected subgroup.

Today I want to talk about how much regression there is to correct for and how well alternative correction procedures work. I'm going to begin, however, with a brief discussion of the regression phenomenon itself.

Discussions of the regression-effect bias dates back at least to Thorndike's 1924 paper in the Journal of Experimental Psychology. Even today, however, it remains a poorly understood phenomenon. I am assuming, therefore, that it will be worthwhile to review some fundamentals.

Figure 1 depicts a normal or Gaussian distribution of test scores. I have drawn some arrows in it to represent the fact that, if the same students were retested with the same instrument or an alternate form of it, most of their scores would change. Some students' retest scores would be higher, and other students' retest scores would be lower. The changes would approximately balance out, however, and the distribution of retest scores would look pretty much the same as the distribution of original-test scores.

The situation becomes substantially different if, on the basis of one set of test scores, we select either a high- or a low-scoring subgroup. If we retest the selected subgroup, the distribution of their retest scores would look quite different from the distribution of their original test scores. The reason is that the movement of scores we see in Figure 1 is allowed to proceed in only one direction across the cutoff score.

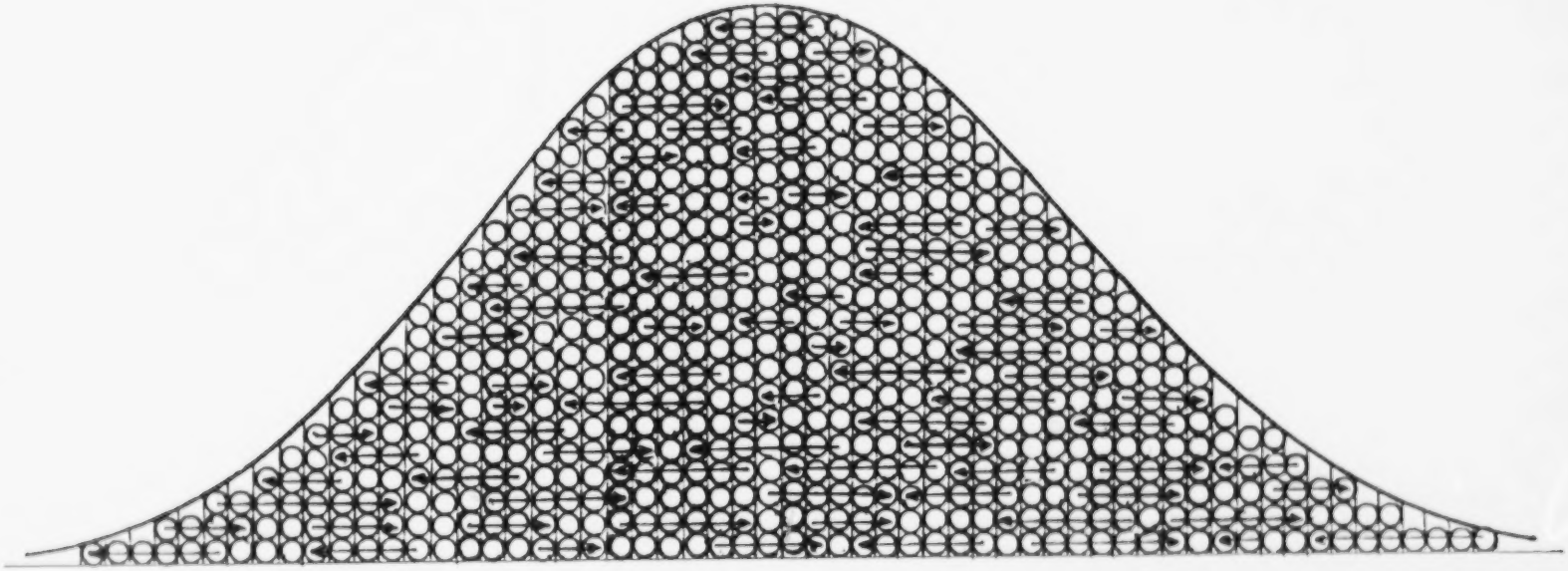

Figure 1. The shape of a normal distribution remains the same from test to retest although the scores of individual students change.

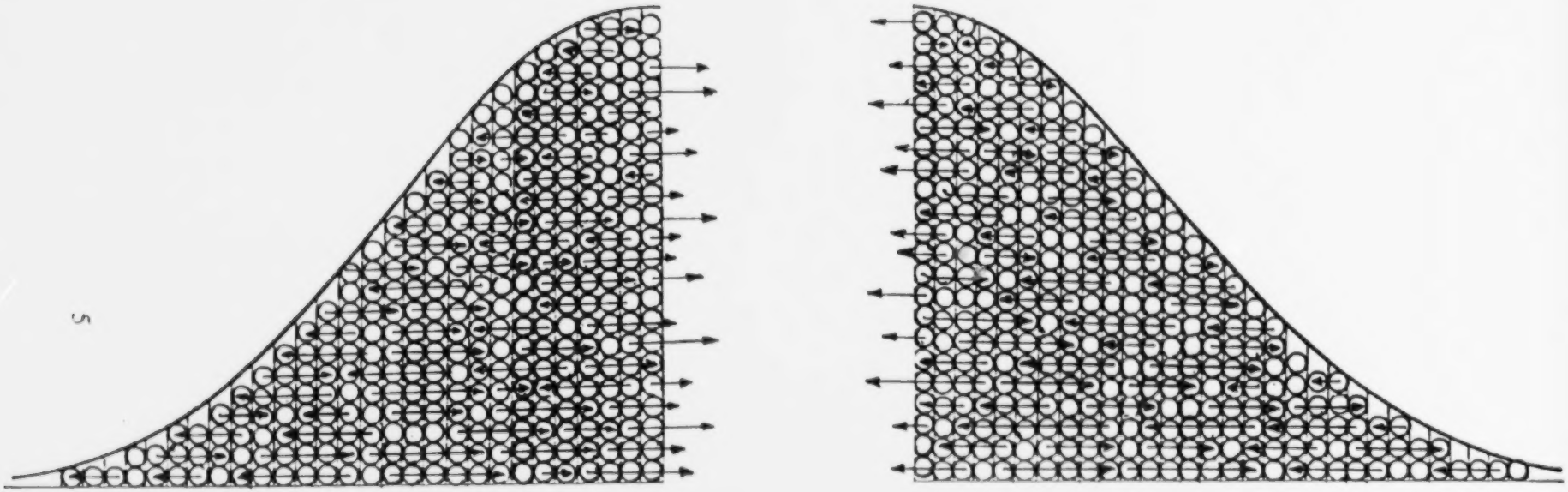

Figure 2. When subgroups are selected, retest scores migrate across the cutoff value.

Scores within the selected subgroup are allowed to cross the cutoff, but scores from the unselected subgroup are not allowed to replace them as they do in the total distribution of scores. This one-way movement of scores across the cutoff value is the sole cause of the regression effect.

The one-way movement of scores across the cutoff value is illustrated in Figure 2. Here the original, intact distribution of scores has been cut approximately in half, and the two halves have been pulled apart. As you can see, some scores from each half cross the cutoff value. If we pushed the two halves back together, the shape of the original distribution would be retained. When we pull them apart, however, each piece begins to grow a new tail. Figure 3 illustrates, in general, the shapes the truncated distributions of original test scores would assume upon retesting.

Scores which were originally on one side of the cutoff value have moved across it. One effect of this movement is to pull the mean of each selected subgroup in the same direction--toward the mean of the total original group. Let me say once more that this migration of scores across the cutoff value is the sole cause of the regression effect.

Now let's look at the regression effect from a more realistic and less theoretical perspective. Table 1 presents three sets of test scores for 100 cases. We can consider the X set of scores as selection test scores, the Y set as pretest scores, and the Z set as posttest scores. Alternatively, we could also consider the X data as selection/pretest scores and the Y and Z data as sets of posttest scores.

All three sets of scores have means of 50 and standard deviations of 20. These values were deliberately chosen to approximate the distribution characteristics of NCEs for a nationally representative sample. Hopefully, that parallelism will provide some experiential basis for readers to use in judging the size of the effects that we will be discussing. A regression effect of, say, 4 points in this data set will be roughly equivalent to a 4 NCE gain in a compensatory reading project, for example.

The data are not real. They have been synthesized. They are not textbook data, however. The bivariate distributions are neither normal nor homoskedastic. They are intended to resemble real data, but we have deliberately avoided building in the kind of ceiling and floor effects that are frequently encountered in situations when test difficulty levels are inappropriate for the students tested.

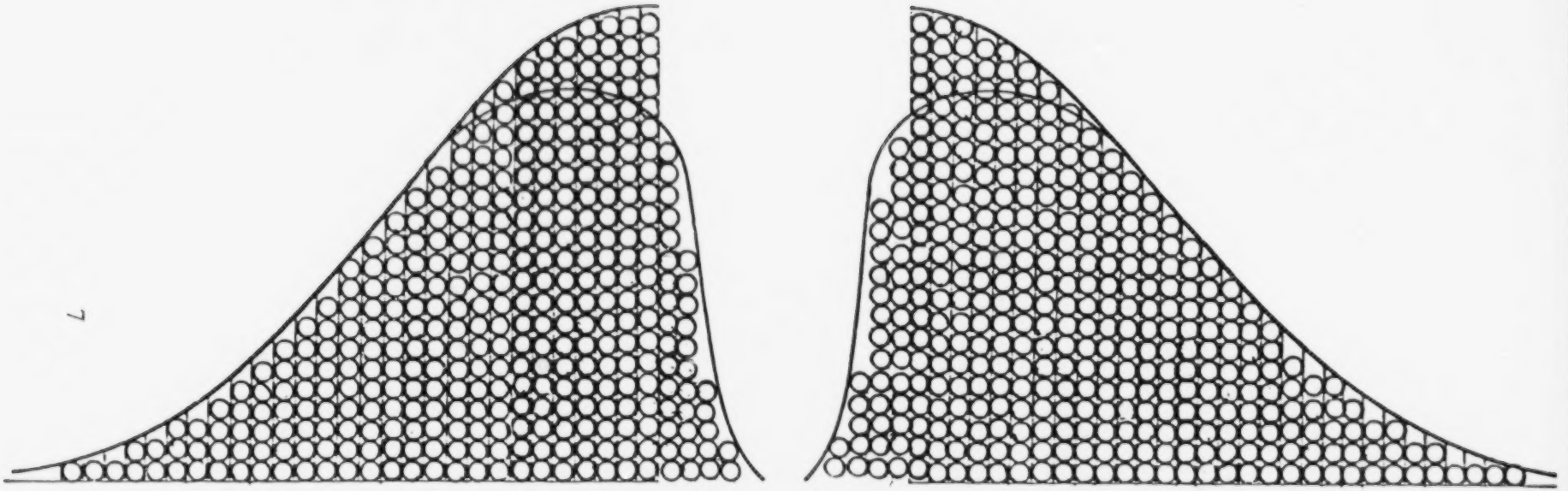

Figure 3. Retest scores that migrate across selection cutoff values produce new tails on truncated distributions.

Table 1

Raw Data

CASE	X	Y	Z
1	4	10	12
2	6	4	8
3	12	6	10
4	14	16	20
5	14	30	38
6	16	28	36
7	20	36	42
8	22	30	26
9	22	50	34
10	22	20	16
11	22	24	20
12	26	14	46
13	26	18	17
14	26	28	63
15	28	54	34
16	28	48	48
17	30	14	28
18	30	50	32
19	30	26	14
20	32	36	49
21	36	36	8
22	36	28	51
23	36	56	34
24	36	42	48
25	36	54	52
26	38	28	49
27	38	42	44
28	38	52	48
29	40	48	28
30	40	62	62
31	40	32	42
32	40	50	50
33	42	22	22
34	42	24	42
35	42	44	32
36	42	50	44
37	42	38	60
38	42	54	49
39	42	64	48
40	44	38	27
41	44	46	30
42	46	56	64
43	48	54	40
44	48	28	48
45	50	38	44
46	50	48	52
47	50	50	55
48	50	52	43
49	50	56	26
50	50	64	54

CASE	X	Y	Z
51	50	54	60
52	50	32	46
53	52	34	34
54	52	50	68
55	52	56	48
56	52	54	56
57	52	52	72
58	52	72	52
59	54	44	50
60	54	54	48
61	54	26	57
62	54	50	64
63	54	70	58
64	54	52	66
65	54	68	54
66	58	34	36
67	58	48	68
68	58	74	62
69	58	46	50
70	60	76	71
71	60	60	51
72	60	56	63
73	62	46	49
74	62	56	32
75	64	66	52
76	64	76	78
77	66	52	50
78	66	62	92
79	66	64	80
80	68	50	61
81	70	54	32
82	70	64	66
83	70	76	74
84	70	86	68
85	72	72	48
86	72	56	66
87	74	54	51
88	74	46	28
89	74	78	86
90	74	80	64
91	76	88	86
92	78	54	69
93	78	78	90
94	80	72	78
95	86	78	83
96	88	70	68
97	88	96	52
98	90	82	87
99	92	94	98
100	96	90	89

The scatter plot shown in Figure 4 has X values along the horizontal axis and Y values along the vertical axis. The X-Y correlation is .800. Figures 5 and 6 present the X-Z and Y-Z scatter plots respectively. The corresponding correlations are .731 and .740.

A few of the data points in the Figure 4 scatter plot fall exactly on the main diagonal of the distribution. These points represent students who got exactly the same scores on the original test and the retest. The scores of all the other students changed from one test to the other--just like the scores moved in the idealized distributions shown in Figures 1, 2, and 3.

In Figure 4, I have circled the scores of four students who scored well below 40 on the first test and well above 40 on the second test. Had 40 been the cutoff score for selection into a compensatory education project, the second-test scores of these students would have migrated across it and, along with the other migrating scores, would have served to raise the mean second-test score above that of the first-test scores for the selected subgroup. Figure 7 depicts the same phenomenon schematically. The shaded portion of the diagram represents students whose scores were below the cutoff on the first test but above the same cutoff on the second test.

To look at the size of the regression effect and the adequacy of correction procedures, I used the data presented in Table 1 and did analyses for all possible cutoff scores. The summary statistics that underly these analyses are presented in Table 2.

There were 36 possible cutoff scores and, for each of them, I calculated the means, standard deviations, and intercorrelations of those scores that fell below the cutoffs. As you can see, all of the Z means are larger than the corresponding X means. All but the first Y mean are also larger than the corresponding X means, and almost all of the Z means are larger than the corresponding Y means.

What I call "actual regression" is the mean of Y minus the mean of X or the mean of Z minus the mean of X. I think you can see in Table 2 that the amount of actual regression is greatest when the mean of the selected subgroup is farthest below the mean of the total group. It approaches zero as the two means approach each other. This is consistent with the prediction equation presented earlier.

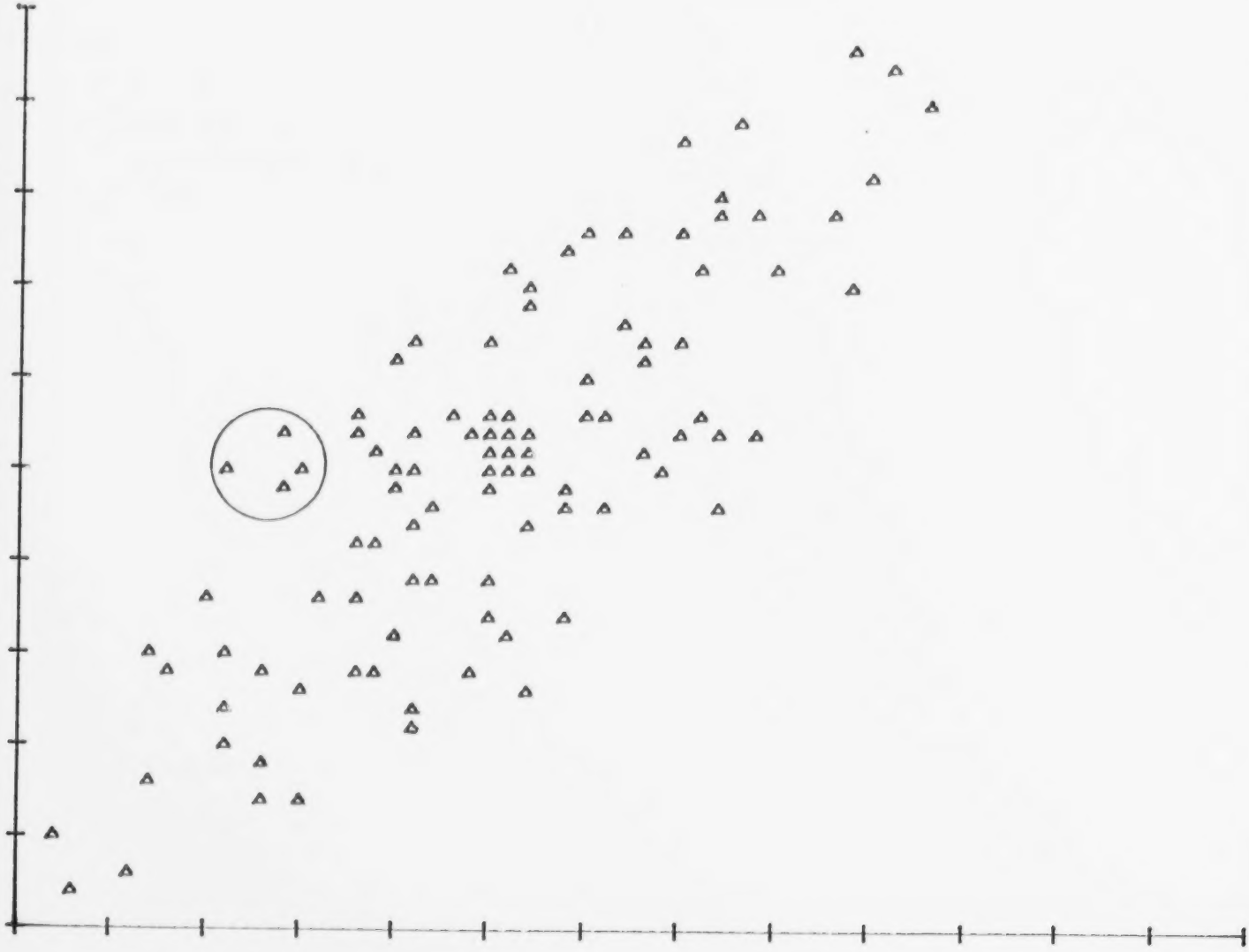

Figure 4. Scatter plot of X and Y scores.

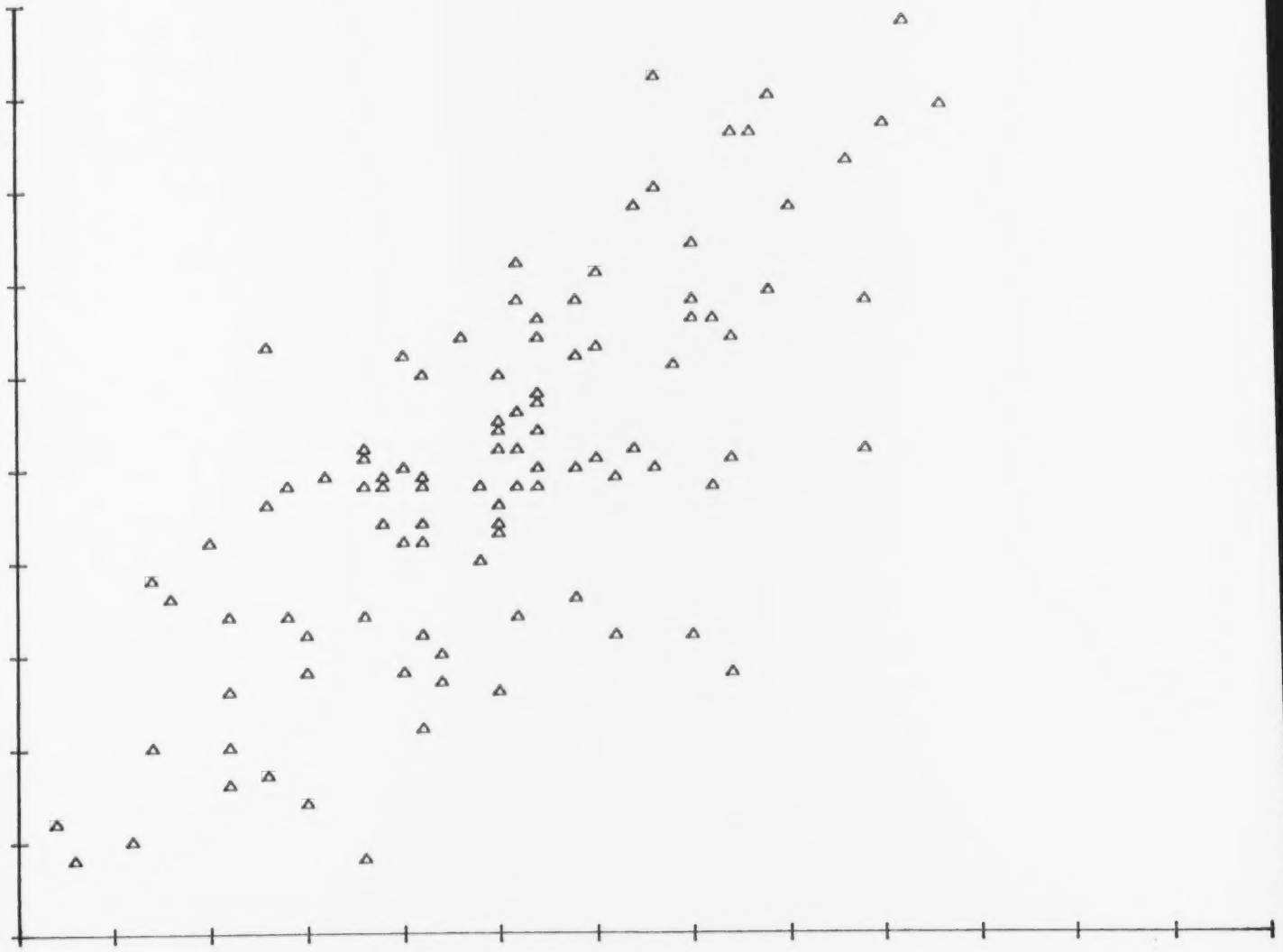

Figure 5. Scatter plot of X and Z scores.

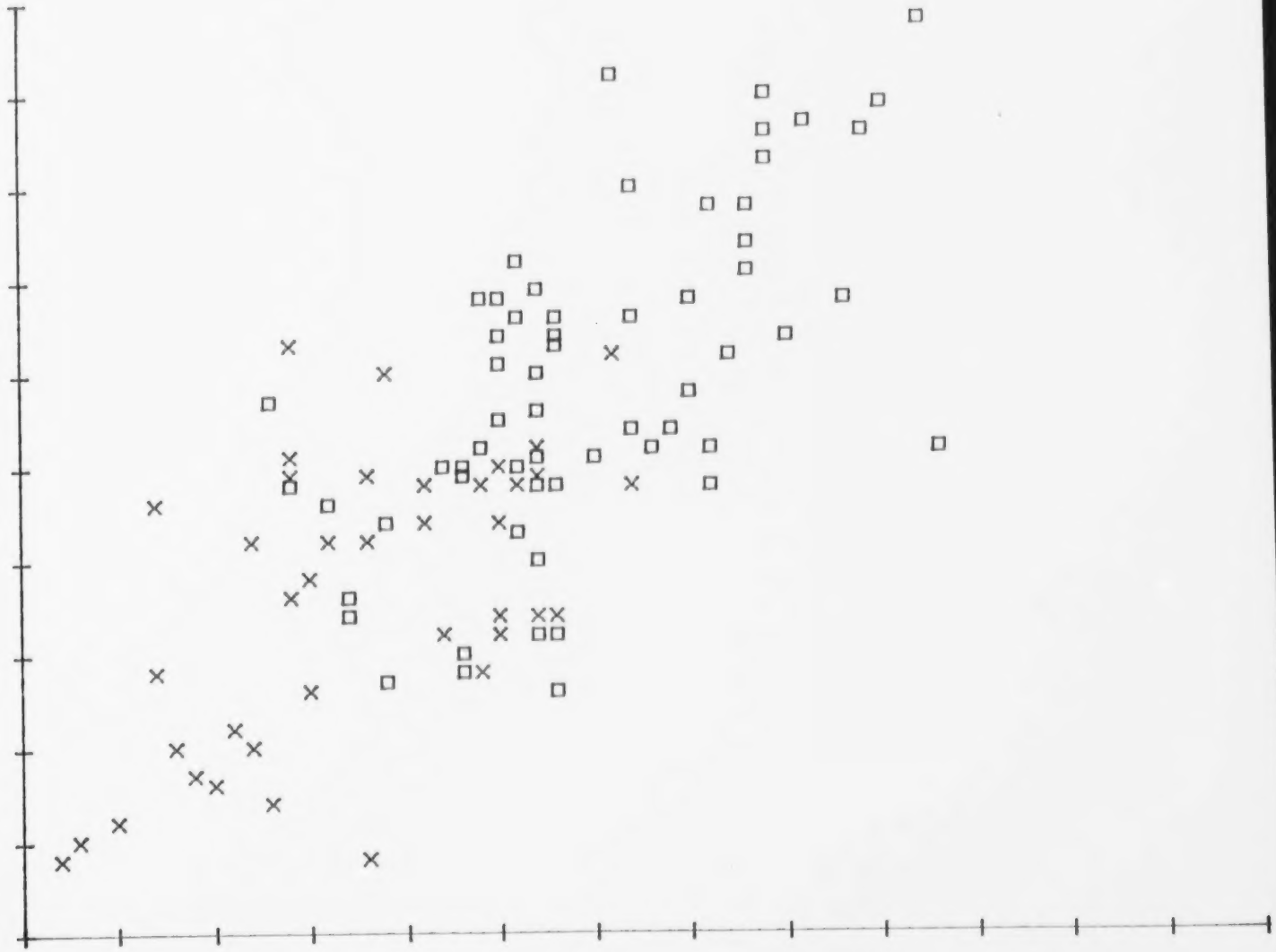

Figure 6. Scatter plot of Y and Z scores. Boxes are YZ scores above X = 39.5 and Xs are YZ scores below X = 39.5

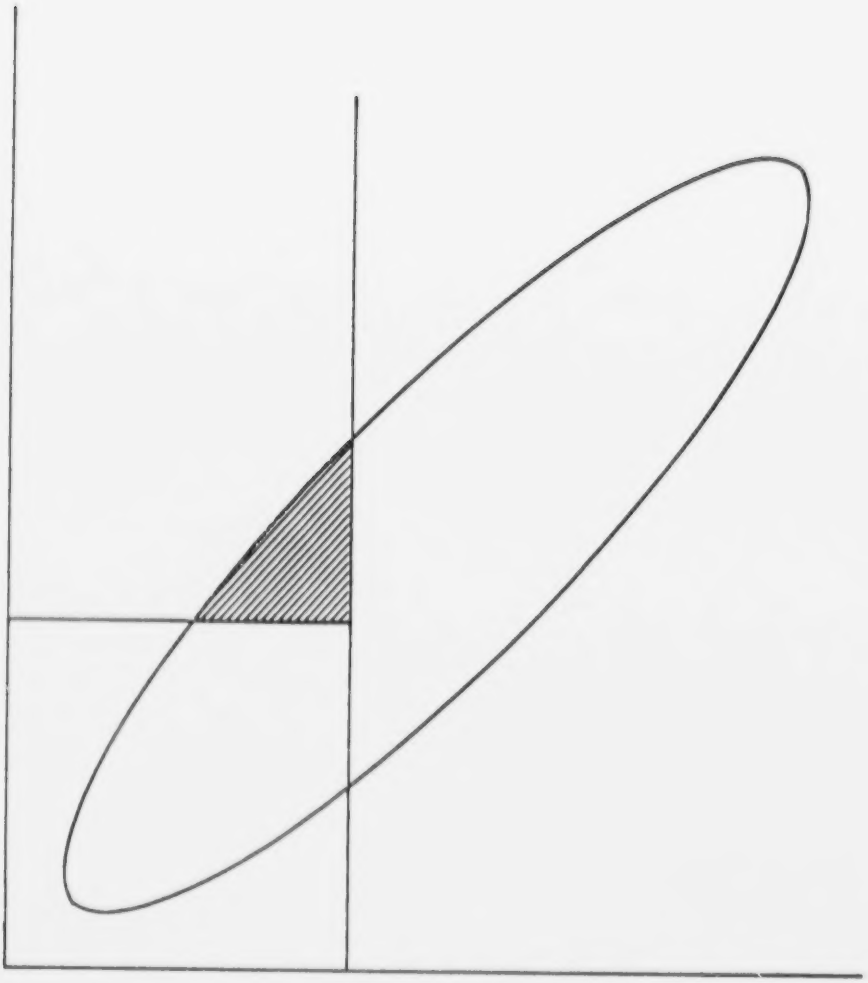

Figure 7. A bivariate distribution showing second-test scores that migrated across the first-test cutoff values.

Table 2

Selected Subgroup Summary Statistics

N	$\bar{X}$	$\bar{Y}$	$\bar{Z}$	SD. X	SD. Y	SD. Z	R XY	R XZ	R YZ
3	7.33	6.67	10.00	3.40	2.49	1.63	-0.491	-0.240	0.982
5	10.00	13.20	17.60	4.20	9.35	10.98	0.612	0.642	0.994
6	11.00	15.67	20.67	4.43	10.16	12.15	0.718	0.743	0.996
7	12.29	18.57	23.71	5.17	11.79	13.50	0.821	0.827	0.995
11	15.82	23.09	23.82	6.24	13.14	11.52	0.730	0.521	0.854
14	18.00	22.43	27.71	6.93	12.03	15.40	0.501	0.568	0.572
16	19.25	26.00	29.38	7.28	14.73	15.26	0.632	0.608	0.585
19	20.95	26.63	28.63	7.74	14.84	14.44	0.546	0.448	0.539
20	21.50	27.10	29.65	7.92	14.61	14.76	0.558	0.499	0.551
25	24.40	30.32	31.44	9.16	15.32	15.56	0.634	0.473	0.513
28	25.86	31.43	33.11	9.62	15.17	15.49	0.637	0.540	0.528
32	27.63	33.50	34.66	10.14	15.68	15.68	0.673	0.563	0.570
39	30.21	35.08	36.05	10.72	15.81	15.30	0.628	0.549	0.575
41	30.88	35.41	35.68	10.87	15.51	15.01	0.626	0.495	0.558
42	31.24	35.90	36.36	10.98	15.64	15.45	0.642	0.523	0.581
44	32.00	36.14	36.70	11.28	15.57	15.20	0.621	0.526	0.564
52	34.77	38.15	38.37	12.24	15.54	15.02	0.646	0.553	0.568
58	36.55	39.69	40.09	12.72	15.80	15.63	0.666	0.592	0.592
65	38.43	41.02	41.88	13.18	16.05	15.77	0.662	0.640	0.593
69	39.57	41.57	42.58	13.59	16.13	15.84	0.649	0.642	0.605
72	40.42	42.50	43.38	13.91	16.51	16.06	0.673	0.662	0.631
74	41.00	42.73	43.30	14.16	16.36	15.91	0.670	0.632	0.618
76	41.61	43.47	43.87	14.45	16.78	16.22	0.692	0.646	0.639
79	42.53	44.08	45.01	14.92	16.77	17.26	0.702	0.670	0.650
80	42.85	44.15	45.21	15.10	16.68	17.25	0.696	0.674	0.650
84	44.14	45.38	45.92	15.83	17.38	17.50	0.722	0.669	0.667
86	44.79	45.81	46.17	16.19	17.45	17.43	0.728	0.666	0.661
90	46.09	46.64	46.67	16.94	17.76	17.75	0.731	0.643	0.676
91	46.42	47.10	47.10	17.13	18.18	18.12	0.741	0.657	0.693
93	47.10	47.51	47.80	17.55	18.28	18.59	0.743	0.678	0.704
94	47.45	47.77	48.12	17.78	18.36	18.75	0.749	0.688	0.710
95	47.85	48.08	48.48	18.12	18.52	18.99	0.757	0.700	0.719
97	48.68	48.80	48.72	18.82	19.08	18.90	0.772	0.689	0.703
98	49.10	49.14	49.11	19.18	19.27	19.19	0.780	0.703	0.713
99	49.54	49.60	49.61	19.56	19.69	19.71	0.791	0.719	0.729
100	50.00	50.00	50.00	20.00	20.00	20.00	0.800	0.731	0.740

You can also see that the amount of actual regression from X to Z is greater than that from X to Y. This, too, is consistent with the prediction equation presented earlier, as the X-Z correlation (.731) is lower than the X-Y correlation (.800).

Table 3 shows these relationships more clearly. Table 3 also includes data on predicted regressions and on the differences between predicted and actual regressions. In columns 3 and 8 of Table 3, the predicted regressions were derived from the formula presented earlier using the actual X-Y and X-Z correlations for the total group (.800 and .731 respectively). The predicted regressions in columns 4 and 9 were derived from the same formula but, instead of using the actual total-group X-Y correlation, they were computed using estimates of the total-group correlation derived from the X-Y correlations within the selected subgroups. The estimated total-group correlation was calculated using the following formula:

$$K_{xy} = \frac{r_{xy}(SD_x/sd_x)}{\sqrt{1 - r^2_{xy} + r^2_{xy}(SD_x/sd_x)^2}}$$

The upper-case letters refer to total-group quantities and the lower-case letters refer to selected-subgroup quantities.

The first thing to note in Table 3 is that the numbers in the columns labeled actual regressions are substantial. If we assume, as is typical of the Chapter 1 program, that students scoring below the 40th percentile of the national norms are selected to receive services, the actual X to Y regressions are about 4.5 points (or NCEs) while the X to Z regressions are somewhat higher. If a project does not correct for these regression biases, the effect-size estimate obtained from implementing a norm-referenced evaluation will be inflated by at lease 4.5 NCEs. The spurious gain component would than be a minimum of four and a half times larger than the national average true gain which has been estimated at approximately 1 NCE. Clearly, it is essential to correct for the regression-effect bias when project participants are selected by virtue of obtaining low pretest scores.

I want to focus attention on that portion of Table 3 between N = 19 and N = 52 simply because this is the range in which students might be selected to participate in compensatory or remedial education programs. Within this range, the average actual regression from X to Y is 5.02 points or NCEs. From X to Z, it is 6.12. The average regression from Y to Z is thus 1.10 points or NCEs.

Table 3

Actual and Predicted Regressions

N	ACTUAL REG. (Y - X)	PRED. REG.1	PRED. REG.2	A - P1	A - P2	ACTUAL REG. (Z - X)	PRED. REG.1	PRED. REG.2	A - P1	A - P2
3	-0.66	8.53	83.52	-9.19	-84.18	2.67	11.48	62.44	-8.81	-59.77
5	3.20	8.00	1.40	-4.80	1.80	7.60	10.76	-0.58	-3.16	8.18
6	4.67	7.80	0.87	-3.13	3.80	9.67	10.49	-0.52	-0.82	10.19
7	6.28	7.54	0.59	-1.26	5.69	11.42	10.14	0.31	1.28	11.11
11	7.27	6.84	1.37	0.43	5.90	8.00	9.19	11.26	-1.19	-3.26
14	4.43	6.40	4.54	-1.97	-0.11	9.71	8.61	0.47	1.10	9.24
16	6.75	6.15	2.67	0.60	4.08	10.13	8.27	3.85	1.86	6.28
19	5.68	5.81	4.07	-0.13	1.61	7.68	7.81	8.92	-0.13	-1.24
20	5.60	5.70	3.94	-0.10	1.66	8.15	7.67	6.79	0.48	1.36
25	5.92	5.12	3.25	0.80	2.67	7.04	6.89	9.49	0.15	-2.45
28	5.57	4.83	3.28	0.74	2.29	7.25	6.49	6.87	0.76	0.38
32	5.87	4.47	2.83	1.40	3.04	7.03	6.02	6.48	1.01	0.55
39	4.87	3.96	3.31	0.91	1.56	5.84	5.32	5.71	0.52	0.13
41	4.53	3.82	3.29	0.71	1.24	4.80	5.14	7.05	-0.34	-2.25
42	4.66	3.75	3.07	0.91	1.59	5.12	5.05	6.41	0.07	-1.29
44	4.14	3.60	3.34	0.54	0.80	4.70	4.84	5.93	-0.14	-1.23
52	3.38	3.05	2.89	0.33	0.49	3.60	4.10	5.00	-0.50	-1.40
58	3.14	2.69	2.50	0.45	0.64	3.54	3.62	3.98	-0.08	-0.44
65	2.59	2.31	2.30	0.28	0.29	3.45	3.11	2.69	0.34	0.76
69	2.00	2.09	2.27	-0.09	-0.27	3.01	2.81	2.38	0.20	0.63
72	2.08	1.92	1.97	0.16	0.11	2.96	2.58	2.13	0.38	0.83
74	1.73	1.80	1.92	-0.07	-0.19	2.30	2.42	2.43	-0.12	-0.13
76	1.86	1.68	1.69	0.18	0.17	2.26	2.26	2.27	.00	-0.01
79	1.55	1.49	1.51	0.06	0.04	2.48	2.01	1.87	0.47	0.61
80	1.30	1.43	1.51	-0.13	-0.21	2.36	1.92	1.75	0.44	0.61
84	1.24	1.17	1.19	0.07	0.05	1.78	1.58	1.65	0.20	0.13
86	1.02	1.04	1.07	-0.02	-0.05	1.38	1.40	1.54	-0.02	-0.16
90	0.55	0.78	0.84	-0.23	-0.29	0.58	1.05	1.34	-0.47	-0.76
91	0.68	0.72	0.75	-0.04	-0.07	0.68	0.96	1.19	-0.28	-0.51
93	0.41	0.58	0.62	-0.17	-0.21	0.70	0.78	0.90	-0.08	-0.20
94	0.32	0.51	0.55	-0.19	-0.23	0.67	0.69	0.78	-0.02	-0.11
95	0.23	0.43	0.46	-0.20	-0.23	0.63	0.58	0.64	0.05	-0.01
97	0.12	0.26	0.28	-0.14	-0.16	0.04	0.36	0.44	-0.32	-0.40
98	0.04	0.18	0.19	-0.14	-0.15	0.01	0.24	0.29	-0.23	-0.28
99	0.06	0.09	0.09	-0.03	-0.03	0.07	0.12	0.14	-0.05	-0.07
100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AVERAGE FROM N = 19 TO N = 52				0.61	1.70				0.19	-0.74

We can think of X as the selection test, Y as the pretest, and Z as the posttest. If this is indeed the case, most (82%) of the regression from selection test to posttest occurs between the selection test and the pretest. This regression is appropriately attributed to random measurement error, and a correction based on the reliability of the test is clearly appropriate.

The regression from Y to Z cannot be attributed to measurement error and must therefore, be regarded as "real" in some sense of that word. Real changes in students' achievement levels that are independent of their pretest status almost certainly account for some of this regression. Part can also be attributed to the fact that the test is measuring somewhat different skills at posttest time than it did at pretest time. Whatever it comes from, it is clear that regression from pre- to posttest did not result from an educational intervention. In our example, the pre-to-posttest regression amounted to 1.1 NCEs, and there was no intervention. Had there been an intervention, its impact would be correctly quantified as pre-to-posttest growth minus the 1.1 NCEs resulting from regression.

Now suppose that the X data are scores on a selection/pretest and that the Z data are scores on a posttest. Clearly, correcting for regression due to measurement error will not be adequate if we wish to assess the impact of an intervention that occurred between X and Z. We must correct both for the regression due to measurement error and for the "real" regression if we wish to have an unbiased effect-size estimate.

"Correcting for regression" involves simply adding the predicted amount of regression to the mean selection/pretest score. In our simulation, where there is no real growth between testings, the corrected mean selection/pretest score should equal the mean posttest score for the selected subgroup and will equal it if the predicted and actual regressions are the same. Put another way, the regression-effect correction will be exactly correct if the predicted regression equals the actual regression. Ideally, then, all values in columns 5, 6, 10, and 11 should be zero.

Both prediction formulas presented above are theoretically correct. It follows, then, that all values in columns 5, 6, 10, and 11 would be zero if our bivariate score distributions were normal and homoskedastic. The observed non-zero differences necessarily result from differences between our data sets and theoretically ideal data sets.

As can be seen in Table 3, the difference scores in columns 5 and 10 are smaller than those in columns 6 and 11. That is because the predicated regressions in columns 5

and 10 were derived from known X-Y correlations while those in columns 6 and 10 involved estimating the correlations as well as the regressions.

In our range of interest--from N = 19 to N = 52--the average score for the more accurate predictions (lumping columns 5 and 10 together) is .40. This means that, if the scores were corrected for regression, the average spurious gain would be less than half an NCE. Uncorrected, the average spurious gain for the same range of cutoff score is 5.57 NCEs.

The standard deviation of the difference scores in the same range of cutoff values is .50. Since the expected mean value is zero, the standard deviation actually gives a better indication of the accuracy of the correction process.

If we look now at the values in columns 6 and 11, the picture is not quite so sanguine. Although the mean difference between actual and predicated regressions is only .48, the standard deviation of those differences is now 1.58. Given an unknown X-Y correlation, this may be a fairly realistic assessment of how accurate our corrections for regression are likely to be with this correction procedure. While more accurate corrections would certainly be desirable, it should be noted that, in all cases, the corrected values are substantially closer to what they should be than the uncorrected values. The average increase in accuracy amounts to 4.11 NCEs for the range of cutoff scores being considered.

The level of accuracy reflected by the figures in columns 5 and 10 can almost certainly never be obtained with real-world data, since the true total-group X-Y (or X-Z) correlations will be unknown. The level of accuracy reflected by the figures in columns 6 and 11, however, can realistically be attained using data collected from the project group.

This simulation sheds no light on the accuracy or potential accuracy of the originally recommended correction procedure which involves using published reliability statistics to estimate the total-group X-Y correlation. If that procedure results in a close approximation of the "real" X-Y correlation, corrections for regression based upon it will approach the accuracy of those presented in column 5 of Table 3. Unless the estimated correlation is "right on", however, the correction procedure will introduce a bias, not just a random error as the more recently recommended procedure does. If the correlation is underestimated, the reliability-based procedure will consistently overcorrect for the regression bias. If the correlation is overestimated, the correction will consistently be too small.

And, of course, if X is the selection/pretest and Z is the posttest, correcting for regression from X to Y won't do the whole job. We will still have that 1.1 NCES of bias remaining. There is no reason, however, why publisher-provided reliability coefficients could not be further adjusted to estimate test-retest reliabilities spanning a 12-month interval. Such estimates would reflect both measurement error and "real" regression and might very well result in more accurate regression-effect corrections than the currently recommended procedure.

REFERENCES

Glass, G. V. (March 8, 1978). Regression effect. Memorandum.

Linn, R. L. (1980). Discussion: Regression toward the mean and the interval between test administrations. New Directions for Testing and Measurement, 8, 83-89.

Linn, R. L. (1982). The validity of the Title I Evaluation and Reporting System. In E.R. Reisner, M.C. Alkin, R. F. Boruch, R. L. Linn, & J. Millman (Eds.), Assessment of the Title I Evaluation and Reporting System. Washington, DC: U.S. Department of Education.

Roberts, A. O. H. (1980). Regression toward the mean and the regression-effect bias. New Directions for Testing and Measurement, 8, 59-82.

Tallmadge, G. K. (1982). An empirical assessment of norm-referenced evaluation methodology. Journal of Educational Measurement, 19 (2), 97-112.

Thorndike, E. L. (1924). The influence of chance imperfections of measures upon the relation of initial score to gain or loss. Journal of Experimental Psychology, 7, 225-232.